DEPARTMENT OF DEFENCE
DEFENCE SCIENCE & TECHNOLOGY ORGANISATION
DSTO

Genetically Optimised Feedforward Neural Networks for Speaker Identification

Richard Price, Jonathan Willmore and William Roberts

DSTO-TN-0203

Genetically Optimised Feedforward Neural Networks for Speaker Identification

Richard Price, Jonathan Willmore & William Roberts

Information Technology Division
Electronics and Surveillance Research Laboratory

DSTO-TN-0203

ABSTRACT

The problem of establishing the identity of a speaker from a given utterance has been conventionally addressed using techniques such as Gaussian Mixture Models (GMMs) that model the characteristics of a known speaker via means and covariances. In this paper we pose the task as a binary classification problem, and whilst in principle any one of a number of classifiers could be applied, this work compares the performance of genetically optimised neural networks versus the conventional approach of GMMs. The test data used in the experiments was the data used for the 1996 National Institute for Standards Technology (NIST) evaluation of speaker identification systems.

DEPARTMENT OF DEFENCE
DEFENCE SCIENCE & TECHNOLOGY ORGANISATION
DSTO

Published by

DSTO
Electronics and Surveillance Research Laboratory
PO Box 1500
Salisbury South Australia 5108 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-010-960
May 1999

APPROVED FOR PUBLIC RELEASE

Genetically Optimised Neural Networks for Speaker Identification

Executive Summary

The problem of establishing the identity of a speaker from a given utterance has been conventionally addressed using Gaussian Mixture Models (GMMs). In this paper we compare the performance of genetically optimised neural networks versus the conventional approach of GMMs.

Authors

Dr. Richard Price
Information Technology Division

Richard Price is a Senior Research Scientist within Information Management & Fusion group, Information Technology Division. His main technical interests are intelligence processing, pattern recognition and data mining.

Jonathan Willmore
Information Technology Division

Jonathan Willmore graduated B.Sc from Adelaide University in 1985, majoring in physics and computer science. He has worked on systems analysis and design for a Naval combat system, and software development for a decision support system for military intelligence analysts, and is currently undertaking research into speaker recognition.

William Roberts
Information Technology Division

William Roberts obtained the B.E. (hons), B.Sc. (hons) and the Ph.D. degrees from the University of Adelaide in 1990, the University of Adelaide in 1992, and the George Mason University, Fairfax, VA, in 1996, respectively. His areas of interest are information theory, detection theory, and statistical signal processing applied to speech.. He began working at DSTO as a vacation student in 1988 and he is currently a Research Scientist in Information Technology Division. In 1998 he was awarded a fellowship from the Japanese Society for the Promotion of Science allowing him to undertake post-doctoral studies at the Tokyo Institute of Technology.

Contents

1. INTRODUCTION 1

2. GENETICALLY OPTIMISED FEEDFORWARD NEURAL NETWORKS 1

3. TRAINING THE MODELS 2

4. TESTING THE MODELS 3

5. RESULTS 3

6. CONCLUSIONS 5

7. REFERENCES 5

APPENDIX 1 7

APPENDIX 2 8

1. Introduction

Automatic speaker recognition refers to the process of recognizing speakers from their voices. This process may be performed by comparing the utterance from a speaker of unknown identity with templates or models of various speakers of interest. The degree of similarity between the models and the utterance is then used to make a decision.

The speaker recognition problem, referring to the general area of recognizing speakers from their voices, may be subdivided into smaller problems. Speaker *verification* is the problem of deciding if an utterance is from a particular speaker or not. Speaker *identification* is the problem of deciding who is speaking in a given utterance.

This paper describes a neural network based speaker recognition system in which A single neural network model is constructed for each speaker of interest. The neural network models are trained using feature vectors known as cepstral coefficients. Further information regarding cepstral coefficients is given in ref [1] .

The system was tested using the evaluation data of the 1996 National Institute of Standards and Technology (NIST) Speaker Evaluation Workshop. Using the data from the NIST evaluation, the neural network system was compared against the conventional Gaussian Mixture Model (GMM) based system [1].

2. Genetically Optimised Feedforward Neural Networks

Each neural network was designed to make a binary decision as to whether or not the test utterance is the target speaker for that model or not. This implies that there needs to be a separate network for each target speaker. Whilst it is feasible for one network to be capable of deciding between multiple targets, for practical reasons that approach was not adopted in order to contain the degree of complexity to a manageable level. Thus, if either one of the target speakers was changed or a new target added, this would require the network to be retrained on all speakers. If there is one network per target speaker, and there is a change in speaker, then a new network needs to be developed for the new speaker, but the other models can be left unchanged.

The feature vector presented to the neural network consisted of the top 20 cepstral coefficients with the output pattern being one of two possibilities. An output pattern 1 0 indicates that this vector is the target speaker for the network, and the pattern 0 1 indicates that the vector is from the corpus of background speakers. For binary classification problems, it is possible to have just the one output node, where a 1 indicates the target and a 0 the background, however better performance is generally reported to be obtained with two outputs. The background corpus comprised of many vectors from a large number of speakers other than the target speaker. This data is required so the neural network can learn the characteristics that distinguish

the target speaker from other speakers. This process is commonly known as competitive or supervised learning. In order not to bias the network, fifty percent of the training data was taken from the target speaker for that neural network model and fifty percent was taken from the background corpus.

The optimal topology of each neural network is generated using genetic algorithms. The topology of a neural network is defined by the number of inputs, the number of hidden layers, the number of nodes on each hidden layer, the transfer function used on each neuron and the learning paradigm adopted. A cost function is created representing the fitness of each trained network. The fitness function is a weighted function combining each network's performance against training data and unseen test data. The intention behind the approach is that when the fitness function is maximised, the network has not only learnt the training data but has also successfully generalised to the unseen test data. The cost function is expressed as a function of the neural networks topology variables. By experimenting over a large range of different networks and selecting those networks that perform better at learning and generalising on unseen data, the cost function is maximised using a genetic algorithm, to obtain the optimal network topology for the given problem and associated data. Further details on the cost function are provided in Appendix 2. A commercially available package that implemented genetically optimised neural networks was adopted for this work. Whilst the application of neural networks to the speaker identification problem is not novel, it is the belief of the authors that the application of genetically optimised neural networks are not reported elsewhere in the literature.

3. Training The Models

The system was developed and tested using the evaluation data of the 1996 National Institute of Standards and Technology (NIST) Speaker Evaluation Workshop. The NIST workshop addresses the text independent, open set, speaker verification problem and comprises specific testing and training conditions designed to exercise systems under real-world conditions [6].

For the purposes of this study only male speakers of the NIST evaluation data set were considered. The models were trained on 2 minutes of speech from 2 different handsets for each of the 21 speakers. (ie. approximately 1 minute of speech from each handset). The testing utterance length was nominally 30 seconds, and the results were split according to the microphone used during testing. The matched results were those where the same microphone was used for both testing and training, whereas the mis-matched results were those for when different microphones were used during testing and training.

For this experiment a background model was not required, as background data was evenly interspersed with target data. Thus for each target model constructed, 50% of the vectors used are actual target vectors and the other 50% are background vectors, where the background data was obtained from the 1996 NIST development corpus. There were no target speakers present in the background data. Feature vectors, (in this case the first twenty cepstral coefficients) were calculated for each waveform

and used in the training and testing of the neural networks. Cepstral coefficients were chosen due to their features of high data reduction, high performance, and robustness to certain channel effects. Neural network models were produced for each of the twenty-one speakers in the corpus, and tested against the test data.

4. Testing The Models

Since the neural networks are trained on a vector by vector basis, the testing is performed in the same manner. Therefore, each test utterance is split up into a sequence of vectors each of which are tested against each neural network in turn. If an utterance consists of T vectors, each network will return T output vectors of length 2 for that utterance. The score for each network on a given utterance is the mean of the first output node taken across all the vectors in the utterance. The closer the mean value is to one, the stronger the belief that the utterance is the target speaker, and the closer the mean value is to zero the stronger the belief the utterance is not the target speaker. When all the utterances have been tested against all the neural network models, the model with the highest mean is identified as the target speaker.

5. Results

Two types of errors are possible, namely, misses and false alarms. A false alarm occurs when an utterance is considered to be a target speaker when it is not, and a miss occurs when an utterance is considered to be a background speaker when it is actually a target speaker. For a given test, the probabilities of these two types of errors may be traded off against each other by varying the decision threshold. The resulting plot is called a Detection Error Trade-off (DET) curve (see Appendix A for details).

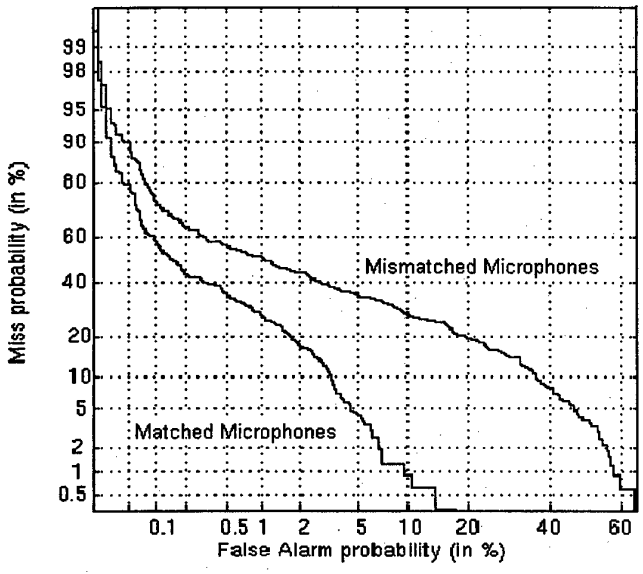

Figure 1: Neural Network DET curve

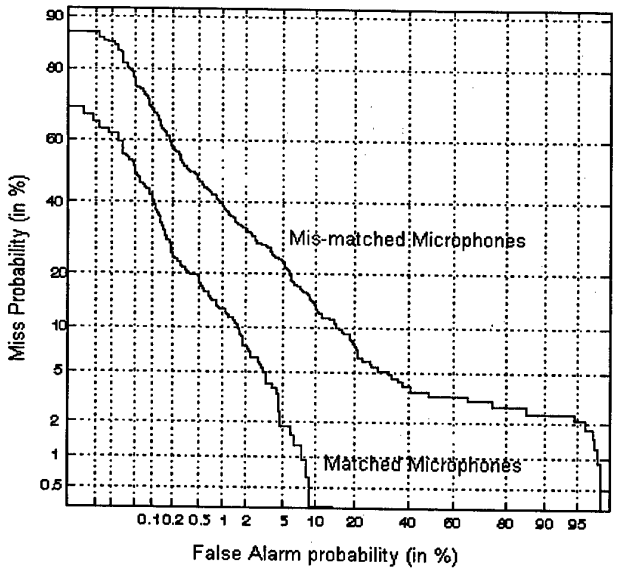

Figure 2: GMM DET curve

Figures 1 and 2 show the performance of the neural network and GMM systems under matched and mis-matched microphone conditions. In order to compare different systems, a single reading is taken from each curve to represent a system's performance. This point is conventionally chosen as the point where the rate of miss is equal to the rate of false alarm, and is commonly referred to as the equal error rate. Clearly, the lower the equal error rate the better the system. The equal error rates (ERR) in fig.1 for the neural network are approximately 5% for matched and 20% for mis-matched microphones respectively. The GMM system [1] obtained equal error rates of 4% and 13% for matched and mis-matched microphones respectively. From these results we can conclude that the two systems perform similarly when the microphone used for the training and test data is the same, however the GMM system outperforms the neural network when different microphones are used. This would imply that the neural network is not generalizing as well as the GMM to changes between the training and test environments.

However whilst equal error rates are convenient for comparing the performance of different systems, they do not necessarily determine the best system for a particular application. In many applications users tend not to operate at the equal error rate. For example, it may be that a high false alarm rate will be tolerated in order to minimise the chance of missing a target speaker. When one compares the performance of the two systems in this scenario for the mis-matched case, the neural network outperforms the GMM. In order to achieve a 0.5% chance of miss, it can be seen from Fig2 that the GMM has a 98% chance of false alarm, whereas it can be seen on Fig 1 the neural network has only a 64% chance of false alarm. However interestingly, for the matched case the opposite is true, with the neural network giving a 14% false alarm rate and the GMM only 9%.

6. Conclusions

An important factor to be considered in this work is the ease at which the neural network solution was obtained. For practitioners from non-speech processing disciplines the task of implementing a Gaussian Mixture Model would be daunting.

This work highlights that if the training and testing environments are similar, then a user-friendly commercially available package can be applied without a significant loss in performance. This is the case for practitioners in fields such as zoology, where there is a growing interest in the classification of animal sounds for purposes of identification of species or individuals. Researchers interested in this problem are turning to speech processing techniques such as GMMs. For this application, where the same microphone could be used for training and testing, the neural network approach discussed here, using commercially available packages, would also appear appropriate.

The results presented indicate both the GMM and neural network systems have similar performance characteristics when the same microphone is used for training and testing. The GMM however, generalises better than the neural network at identifying the speaker when different microphones are used. The importance of this result would depend on the particular application being considered.

Further work in the application of neural networks to speaker identification should address the following issues. An investigation into the optimum number of cepstral coefficients input into the neural network. The first twenty were chosen for this experiment due to work performed by Roberts [1] which confirmed this was optimal for GMMs, however this need not be the case for neural networks. Indeed it may well be that the use of cepstral coefficients discards information that a neural network might find beneficial when classifying speakers, and that training on the raw waveform may well prove successful. As noted previously, the novelty in the work presented here is the application of genetically optimised neural networks to the speaker identification problem.

7. References

1. W.J.J. Roberts. "Automatic speaker recognition using statistical models", Technical Report DSTO-RR-0131, Defence Science and Technology Organisation, Jun 1998

2. D.A. Reynolds. "Speaker identification and verification using Gaussian Mixture speaker models", Speech Communication 17(1995), pp.91-108

3. D.A. Reynolds. "Robust text-independent speaker identification using Gaussian Mixture speaker models", IEEE Transactions on Speech and Audio Processing, Vol. 3, No.1, January 1995

4. J. Oglesby and J.S. Mason, " Optimization of neural models for speaker identification," Proc. ICASSP-90, PP261-264, 1990

5. Y. Konig, N. Morgan and C. Chandra, "GDNN A gender dependent neural network for continuous speech recognition", International Computer Science Institute Technical Report TR-91-071, December 1991

6. National Institute of Standards Technology, "March 1996 NIST speaker recognition workshop notebook," *NIST administered speaker recognition evaluation on the Switchboard corpus, Mar. 1996.*

7. Naranjo M. Delorme JM. Tuffelli D. "Speaker recognition based on neural networks ",Information Systems Analysis and Synthesis. Proceedings of the International conference on Information Systems Analysis and Synthesis. ISAS'96. Int. Inst. Inf. Syst. 1996, pp.147-53. Orlando, FL, USA.

8. Vieira K. Wilamowski B. Kubichek R." Speaker identification based on a modified Kohonen network.", 1997 IEEE International Conference on Neural Networks. Proceedings (Cat. No.97CH36109). IEEE. Part vol.4, 1997, pp.2103-6 vol.4. New York, NY, USA.

Appendix 1

Appendix - Detection Error Tradeoff Curves

A detection error tradeoff (DET) curve is often used as a means of demonstrating the performance of a speaker verification system. Used in this manner, it is a plot of the false alarm probability versus the miss or detection probability.

Speaker verification constitutes a binary detection problem in which case a decision is made as to whether a hypothesis H1 (i.e. the utterance is a target speaker) is correct, or a hypothesis H2 (i.e. the utterance is a background speaker) is correct.

The hypothesis test in speaker verification using gaussian mixture models is the likelihood ratio test. In this test, the ratio (being the score obtained by the normalization of the probability of the utterance given the target speaker model, to the probability of the utterance given the background speaker model) is compared to a threshold value. When using neural networks a score between zero and one is produced by the network for each of the vectors in the utterance, indicating the strength of belief that vector was spoken by the target. An overall score is computed for the whole utterance by calculating the mean overall the scores for each of the vectors in the utterance. If the mean exceeds 0.5 the conclusion is that the target speaker was speaking and vice versa.

If the ratio exceeds the threshold, then hypothesis H1 is marked as true, else H2 is marked as true. Two types of errors can be made. That of concluding that H1 is correct when it is actually H2, and that of concluding that H2 is correct when it is actually H1. The former is referred to as a false alarm and the latter as a miss. This ratio test is performed for each utterance in a test corpus for a range of threshold values. The number of misses and false alarms are summed and the percentage values are calculated for this threshold. The threshold values are determined by sorting the scores and using one of these scores. Thus for a range of thresholds, a graph can be plotted indicating a miss rate versus false alarm rate for each of these thresholds.

Appendix 2

The penalty function maximised by the genetic algorithm is a linear combination of the percentage of records successfully learnt by a particular neural network on both the training and testing data.

The function has the form Max F = A*train% + B*test% where A and B are user specified, and train% and test% are the percentage of records successfully learnt by the neural network for the training and test data respectively. The value of F is known as the net fitness. Clearly, a high F value indicates that the network has not only learnt the training data but has also successfully generalised to the unseen test data. The choice of A and B indicate the degree of relative importance the user wishes to place upon the training and testing components. For the results reported, these values were both set equal to 1.0, indicating an equal priority between training and testing.

Genetically Optimised Feedforward Neural Networks for Speaker Identification

(Richard Price, Jonathan Willmore & William Roberts)

(DSTO-TN-0203)

DISTRIBUTION LIST

	Number of Copies
AUSTRALIA	
DEFENCE ORGANISATION	
Task sponsor:	
DSD (Dr Ian Doherty)	1
S&T Program	
Chief Defence Scientist)	
FAS Science Policy)	1 shared copy
AS Science Corporate Management)	
Director General Science Policy Development	1
Counsellor, Defence Science, London	Doc Control Sheet
Counsellor, Defence Science, Washington	Doc Control Sheet
Scientific Adviser to MRDC Thailand	Doc Control Sheet
Scientific Adviser - Policy and Command	1
Navy Scientific Adviser	1 copy of Doc Control Sheet and 1 distribution list
Scientific Adviser - Army	Doc Control Sheet and 1 distribution list
Air Force Scientific Adviser	1
Director Trials	1
Aeronautical & Maritime Research Laboratory	
Director	1
Electronics and Surveillance Research Laboratory	
Director	1
Chief Information Technology Division	1
Research Leader Command & Control and Intelligence Systems	1
Research Leader Military Computing Systems	1
Research Leader Command, Control and Communications	1
Head, Information Warfare Studies Group	Doc Control Sheet
Head, Software Systems Engineering Group	Doc Control Sheet
Head, Year 2000 Project	Doc Control Sheet
Head, Trusted Computer Systems Group	Doc Control Sheet
Head, Advanced Computer Capabilities Group	Doc Control Sheet
Head, Systems Simulation and Assessment Group	Doc Control Sheet
Head, C3I Operational Analysis Group	Doc Control Sheet
Head, Information Management and Fusion Group	1
Head, Human Systems Integration Group	Doc Control Sheet
Head, C2 Australian Theatre	1

Head, Information Architectures Group	1
Head, Distributed Systems Group	Doc Control Sheet
Head C3I Systems Concepts Group	1
Head, Organisational Change Group	Doc Control Sheet
Author	1
Publications and Publicity Officer, ITD/EOITD	1 shared copy
DSTO Library and Archives	
Library Fishermens Bend	1
Library Maribyrnong	1
Library Salisbury	2
Australian Archives	1
Library, MOD, Pyrmont	Doc Control Sheet
US Defense Technical Information Center	2
UK Defence Research Information Centre	2
Canada Defence Scientific Information Service	1
NZ Defence Information Centre	1
National Library of Australia	1
Capability Development Division	
Director General Maritime Development	Doc Control Sheet
Director General Land Development	Doc Control Sheet
Director General C3I Development	Doc Control Sheet
Director General Aerospace Development	Doc Control Sheet
Navy	
SO (Science), Director of Naval Warfare, Maritime Headquarters Annex, Garden Island, NSW 2000	Doc Control Sheet
Army	
ABCA Standardisation Officer, Puckapunyal	4
Intelligence Program	
DGSTA Defence Intelligence Organisation	1
Corporate Support Program (libraries)	
OIC TRS Defence Regional Library, Canberra	1
Universities and Colleges	
Australian Defence Force Academy	1
Library	1
Head of Aerospace and Mechanical Engineering	1
Senior Librarian, Hargrave Library, Monash University	1
Librarian, Flinders University	1
Other Organisations	
NASA (Canberra)	1
AGPS	1
State Library of South Australia	1
Parliamentary Library, South Australia	1

OUTSIDE AUSTRALIA

Abstracting and Information Organisations	
Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Scientific Abstracts US	1
Documents Librarian, The Center for Research Libraries, US	1
Information Exchange Agreement Partners	
Acquisitions Unit, Science Reference and Information Service, UK	1
Library - Exchange Desk, National Institute of Standards and Technology, US	1
SPARES	5
Total number of copies:	**52**

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Genetically Optimised Feedforward Neural Networks for Speaker Identification

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Richard Price, Jonathan Willmore & William Roberts

5. CORPORATE AUTHOR

Electronics and Surveillance Research Laboratory
PO Box 1500
Salisbury SA 5108 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TN-0203	AR-010-960	Technical Note	May 1999

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
N/A	JNT 97/011	DSD	16	8

13. DOWNGRADING/DELIMITING INSTRUCTIONS

N/A

14. RELEASE AUTHORITY

Chief, Information Technology Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE CENTRE, DIS NETWORK OFFICE, DEPT OF DEFENCE, CAMPBELL PARK OFFICES, CANBERRA ACT 2600

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CASUAL ANNOUNCEMENT Yes

18. DEFTEST DESCRIPTORS

Neural networks
Speech recognition

19. ABSTRACT

The problem of establishing the identity of a speaker from a given utterance has been conventionally addressed using techniques such as Gaussian Mixture Models (GMMs) that model the characteristics of a known speaker via means and covariances. In this paper we pose the task as a binary classification problem, and whilst in principle any one of a number of classifiers could be applied, this work compares the performance of genetically optimised neural networks versus the conventional approach of GMMs. The test data used in the experiments was the data used for the 1996 National Institute for Standards Technology (NIST) evaluation of speaker identification systems.